Large-scale simultaneous testing with hypergeometric inverted-beta priors

Nicholas G. Polson
Booth School of Business
University of Chicago

James G. Scott
McCombs School of Business
University of Texas at Austin

Original version: August 2009
Revised: September 2010

Abstract

We develop a new class of distributions for use in large-scale simultaneous testing. These priors are based on hypergeometric inverted-beta priors, and have two main attractive features: heavy tails, and computational tractability. The family is a four-parameter generalization of the normal/inverted-beta prior, and is the natural conjugate prior for a shrinkage coefficients in a hierarchical normal model. Our results emphasize the usefulness of these of heavy-tailed priors in large multiple-testing problems, as they have mild rate of tail decay in the marginal likelihood $m(y)$—a property long recognized to be important in testing.

We apply our proposed methodology by testing historical patterns of ROA (return on assets) for a cohort of 11,298 publicly traded firms across 93 countries. Our goal is to determine which firms, if any, have systematically outperformed their peer groups over the past 45 years. We find evidence that demonstrably superior performance is quite rare. We compare our findings with the popular literature on corporate success. By our reckoning, these books appear to be studying a sample wherein the majority of firms have performance profiles that are statistically indistinguishable from random chance. These conclusions are consistent with other recent studies on the subject (e.g. Henderson et al., 2009).

Keywords: corporate benchmarking; inverted-beta prior; multiple testing; normal scale mixtures; sparsity

1 Introduction

1.1 Motivating example: historical corporate performance

Understanding the reasons why some firms thrive and others fail is one of the primary goals of research in strategic management. Studies that examine successful companies to uncover some "secret recipe" for success are very popular, both in the academic and popular literature.

Before the search for special causes can begin, however, success must be quantified and benchmarked. In this paper, we use a common metric called "return on assets" (ROA), which gives investors some notion of how effectively a company uses its available funds to produce income. This is fundamentally different from a market-based measure like stock returns, and is much more stable across time. The higher a firm's ROA, the better.

In this paper, we use a Bayesian model for multiple testing to compare publicly traded companies against their peers using historical ROA data. The full data set is quite large: 645,456 records from 53,038 companies in 93 different countries, spanning the years 1966–2008. We restrict attention to the 11,298 firms for which at least 10 years of data are available.

Given the large number of tests being conducted, and the frequentist leanings of the management-theory community, maintaining control over false positives is crucial. Yet having access to the posterior distribution of effect sizes can greatly inform follow-up case studies of individual firms, and is only possible under a fully Bayesian model. This applied context makes a combined Bayes/frequentist approach especially appealing.

1.2 Large-scale simultaneous testing

Large-scale simultaneous testing seeks to uncover lower-dimensional signals from high-dimensional data. For example, researchers who use microarrays have long been interested in the problem of multiplicity adjustment, where "adjustment" can be understood in sense of adjusting one's tolerance for surprise as the set of potentially surprising events grows large. The same issue arises in all modern high-throughput experiments; other examples include functional magnetic-resonance imaging, environmental sensor networks, combinatorial chemistry, and proteomics. Too many Type-I errors will mean too many expensive wild-goose chases. Hence the case for a testing procedure that displays good frequentist properties is very compelling.

But so too is the case for a model-based Bayesian procedure. These experiments may involve thousands of separate tests, and such a large volume of data often allows the distributional properties of "signals" and "noise" to be characterized quite precisely.

This paper considers a new version of the two-groups multiple-testing model, where we observe data y_i for $i = (1, \ldots, p)$ according to a hierarchical model:

$$\begin{aligned} (y_i \mid \beta_i, \sigma^2) &\sim \mathrm{N}(\beta_i, \sigma^2) \\ (\beta_i \mid w, \theta) &\sim w \cdot g(\beta_i \mid \theta) + (1 - w) \cdot \delta_0 \\ w &\sim p(w)\,, \end{aligned}$$

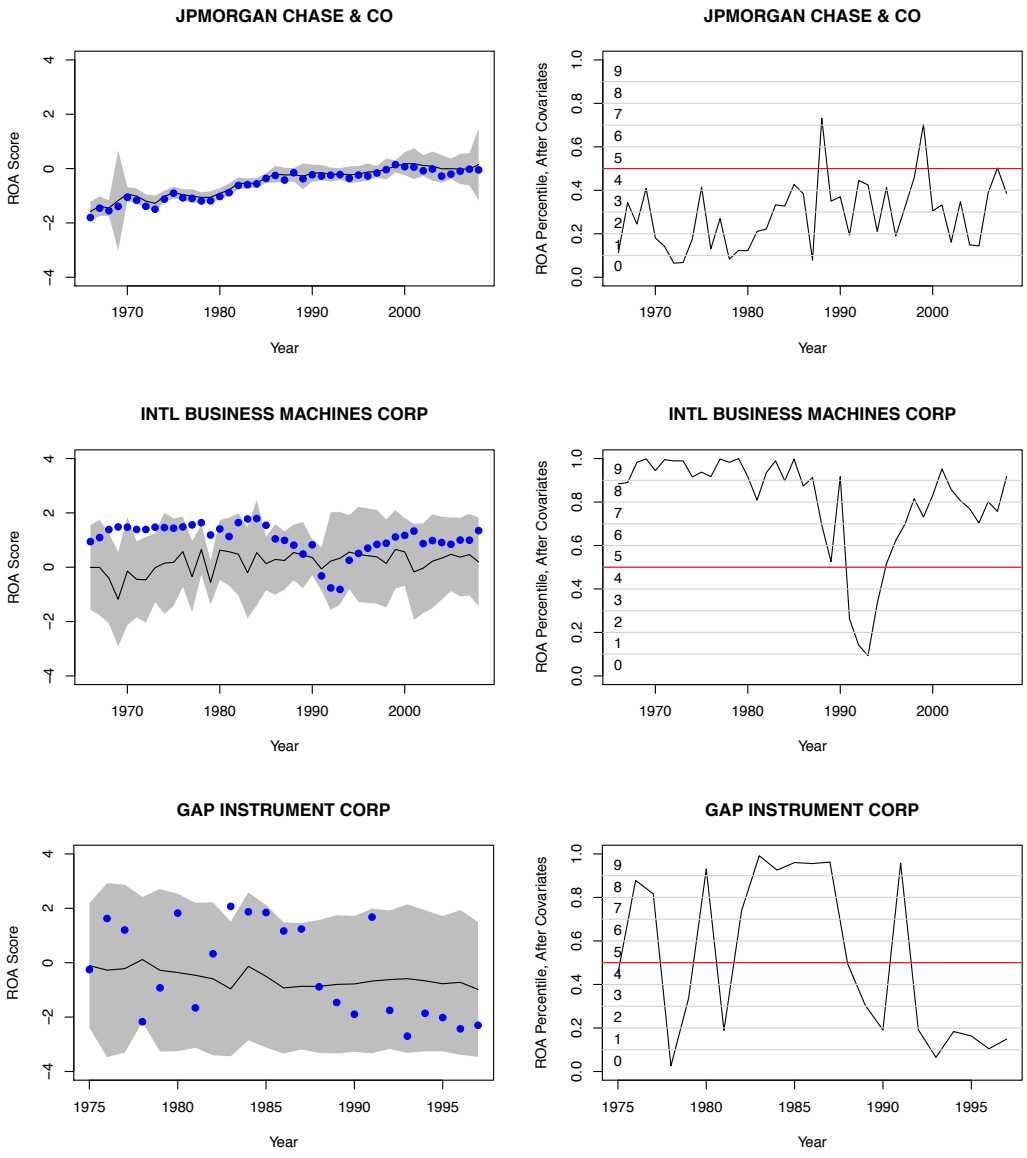

Figure 1: Left: the actual performance of three firms (blue dots), superimposed on the benchmark distribution estimated from the Bayesian regression-tree model (black line and grey area, showing the posterior mean and 95% predictive interval of expected performance). Right: these same firms placed on a common $(0, 1)$ scale of benchmarked performance.

a mixture of a Dirac measure at zero and an alternative model g that is absolutely continuous with respect to Lebesgue measure. (The alternative model g has hyperparameter θ, presumably also given a prior.) The most attractive feature of this model is that it automatically adjusts for multiplicity, without the need for ad-hoc regularization. This is because inference for the β_i's will involve the posterior for common mixing fraction, $p(w \mid \mathbf{y})$. If one tests many noise observations in the presence of a few signals, then our estimate of w will be small, making it more difficult for all the observations to overcome the prior belief in their irrelevance. This exerts a powerful form of control over false positives.

To handle the multiple-testing problem, we introduce a family of distributions g based on normal variance mixtures, where the mixing distribution is a hypergeometric inverted-beta (HIB) prior, which we will soon define:

$$\begin{aligned}(\beta_i \mid \lambda_i^2, \gamma_i = 1) &\sim \mathrm{N}(0, \sigma^2\lambda_i^2) \\ \lambda_i^2 &\sim \mathrm{HIB}(a, b, \tau, s)\,,\end{aligned}$$

where the indicator $\gamma_i = 1$ if β_i is nonzero, and zero otherwise. We approach these priors from a unified Bayesian/frequentist perspective, using them to compute not only posterior distributions, but also false-discovery rates, or FDR (Benjamini and Hochberg, 1995). We also study the behavior of the posterior mean, which is competitive with existing gold-standard methods (e.g. Johnstone and Silverman, 2004) under squared-error loss.

In both our simulation studies (Section 3) and data analysis (Section 4), we focus on three key features of our approach:

1. The hypergeometric inverted-beta scale mixtures form an especially flexible class of symmetric, unimodal densities and can accomodate a wide range of tail behavior and behavior near the centering parameter. This class simultaneously generalizes the robust priors of Strawderman (1971) and Berger (1980), the normal-exponential-gamma prior of Griffin and Brown (2005), and the horseshoe prior of Carvalho et al. (2010). The ability of our class to model heavy-tailed distributions with minimal computational fuss is of particular relevance in testing problems (see, for example, Section 5.2 of Jeffreys, 1961).

2. Our class of priors allows very easy computation of a wide array of important Bayesian and frequentist quantities. This includes posterior means, variances, and higher-order moments; posterior null probabilities for individual observations; the score function; false-discovery rates; and local false-discovery rates (Efron, 2008). The ease with which these quantities can be computed all relates to the analytical tractability of the marginal likelihood function $m(y)$, whose importance we describe in Section 1.4. Appendix A provides all the details.

3. Our approach yields testing error rates that are competitive with existing cutting-edge methods. At the same time, it also retains the advantages of a fully Bayesian procedure, in that one has access to the joint posterior distribution of all parameters.

Many of the technical details characterizing the behavior of the basic mixture model can be found in Scott and Berger (2006) and Bogdan et al. (2008a). These authors assume

that the nonzero means follow a normal distribution, an assumption we will generalize in this paper. Do et al. (2005) also provide an interesting variation wherein the nonzero means are modeled nonparametrically using Dirichlet processes

The same issues arise in empirical-Bayes analysis. See, for example, Johnstone and Silverman (2004), Abramovich et al. (2006), and Dahl and Newton (2007). Additionally, Muller et al. (2006), Bogdan et al. (2008b), and Park and Ghosh (2010) all describe the relationship between Bayesian multiple testing and classical approaches that control the false-discovery rate.

1.3 Data pre-processing

Before applying our method, we pre-processed the data as follows. Let y_{it} be the raw data point for company i in year t. Using Bayesian treed-regression software (Gramacy and Lee, 2008), we estimated a mean m_{it} and a standard deviation s_{it}, representing the expected distribution of performance for other firms in company i's peer group. It is necessary to benchmark raw performance numbers because certain features of a company make it intrinsically easier or harder to earn a high ROA. These same facts may also entail different levels of volatility in performance. Such differences, moreover, may be completely unrelated to the differences in managerial talent or performance that we are hoping to detect. As covariates, we used a company's industry, size, leverage, country of operation, and market share.

We then computed a z-score $z_{it} = (y_{it} - m_{it})/s_{it}$ for each company-year data point. These were averaged to form a composite score for each company, defined as $z_i = \bar{z}_i\sqrt{n_i}$, where n_i is the number of observations (ranging from 10 to 43). If every company-year data point were a random draw from a normal distribution with the benchmarked mean and standard deviation, these composite z-scores would be draws from a standard normal distribution.

The regression-tree approach allows us to account for the highly nonlinear, conditionally heteroskedastic relationships present in the data. Figure 1 shows three example firms. The left-hand plots show the actual performance, along with the "benchmark distribution"—that is, the mean and standard deviation of that year's expected performance, given firm-level covariates. The right-hand plots show the performance with respect to the benchmark distribution, all on a common normal-CDF scale. These scores are the raw inputs to our multiple-testing approach.

1.4 The importance of the marginal likelihood function

Many common Bayesian and frequentist treatments of the multiple-testing problem can be understood through the marginal likelihood functions

$$\begin{aligned} m_0(y \mid \sigma^2) &= \mathrm{N}(y \mid 0, \sigma^2) \\ m_1(y \mid \theta) &= \int_{\mathbb{R}} \mathrm{N}(y_i \mid \beta_i, \sigma^2)\ g(\beta_i \mid \theta)\ \mathrm{d}\beta_i \\ m(y \mid \theta, \sigma^2) &= w \cdot m_1(y) + (1 - w) \cdot m_0(y)\,. \end{aligned}$$

First, following Efron (2008), the local FDR and the posterior probability of y_i being noise are given by the same expression:

$$fdr(y) = P(\beta_i = 0 \mid y, \sigma^2, \theta) = \frac{(1-w) \cdot m_0(y)}{m(y)},$$

Furthermore, if we let $F_0(y) = \int_{-\infty}^{y} m_0(u)\mathrm{d}u$, $F_1(y) = \int_{-\infty}^{y} m_1(u)\mathrm{d}u$, and $F(y) = w \cdot F_1(y) + (1-w) \cdot F_0(y)$, then the FDR is the tail area

$$FDR(y) = \frac{(1-w) \cdot F_0(y)}{F(y)}.$$

Secondly, the marginal likelihood function also arises in Masreliez's classic representation of the posterior mean. This gives an explicit expression for the Bayes estimator for β_i under squared-error loss (assuming that $\gamma_i = 1$):

$$\mathrm{E}(\beta_i \mid y, \gamma_i = 1) = y_i + \frac{d}{dy_i} \ln m_1(y_i),$$

versions of which appear in Masreliez (1975), Polson (1991), Pericchi and Smith (1992), and Carvalho et al. (2010). The choice of alternative model $g(\beta_i \mid \theta)$ is crucial, insofar as it helps to determine $m_1(y)$.

At the same time, the prior should have desirable statistical properties, with flat tails being a particularly important feature. The use of heavy-tailed priors for constructing robust shrinkage estimators has a long history, with prominent examples to be found in Strawderman (1971) and Berger (1980). Jeffreys, meanwhile, observed as early as 1939 that heavy-tailed priors play an important role in Bayesian hypothesis testing (see Jeffreys, 1961, a later edition). His arguments have been recapitulated in the context of linear models by Zellner and Siow (1980) and, more recently, Liang et al. (2008).

The practical issue in both problems is roughly the same: that heavy-tailed priors lead to a desirably mild rate of tail decay in the marginal likelihood $m(\mathbf{y})$, but that very few known priors are both heavy-tailed and analytically tractable. Any prior that possesses both properties, as our proposed family has to potential to do with certain hyperparameter choices, is therefore of great potential interest to Bayesians and non-Bayesians alike.

2 The proposed family of priors

2.1 Connection with classical shrinkage rules

Our new class of priors has its genesis in the large body of work on classical shrinkage rules, where a multivariate normal prior $\boldsymbol{\beta} \sim \mathrm{N}(0, \lambda^2 I)$ is assumed, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$. Many common estimators for this problem, both Bayesian and non-Bayesian, are of the form $\hat{\boldsymbol{\beta}}(\mathbf{y}) = \{1 - g(Z)\}\mathbf{y}$ for $Z = \|\mathbf{y}\|^2$ (e.g. James and Stein, 1961; Strawderman, 1971; Stein, 1981; Fourdrinier et al., 1998). The central issue is how to identify "nice" functions $g(Z)$, and how to understand priors for global variance components in terms of

the behavior of the estimators they yield.

The constraint to rationality—that is, the requirement that there exists a prior $p(\kappa)$ such that, for all Z, $g(Z) = E(\kappa|Z)$ under the posterior $p(\kappa \mid Z)$—rules out a wide class of potential estimators. The function $g(Z)$ cannot, for example, be a polynomial of order two or greater. Indeed, the functional form of a $g(Z)$ that respects admissibility will typically be quite complicated.

It is natural to look in the class of estimators where $g(Z) = p(Z)/q(Z)$, a ratio of power-series expansions. One can construct such a $g(Z)$ by assuming that $(\boldsymbol{\beta} \mid \lambda^2) \sim \mathrm{N}(0, \lambda^2 I)$, and then defining $\hat{\boldsymbol{\beta}}(\lambda^2) = \mathrm{E}(\boldsymbol{\beta} \mid \lambda^2, \mathbf{y})$. After removing the dependence upon λ^2 by marginalizing, this leads to

$$\hat{\boldsymbol{\beta}} = E_{\lambda^2|\mathbf{y}}\{\hat{\boldsymbol{\beta}}(\lambda^2)\} = \{1 - E(\kappa \mid Z)\}\mathbf{y}\,,$$

recalling that $\kappa = 1/(1+\lambda^2)$. We can therefore identify $g(Z)$ with $E(\kappa \mid Z)$, the posterior expectation of κ, given Z.

One can define a class of priors for κ indexed by (a, b, τ, s), which we call the hypergeometric inverted-beta class, such that

$$g(Z) = E(\kappa|Z) = \frac{a + p/2}{a + b + p/2} \frac{\Phi_1(b, 1; a + b + p/2 + 1; s + Z/2, 1 - 1/\tau^2)}{\Phi_1(b, 1; a + b + p/2; s + Z/2, 1 - 1/\tau^2)}\,, \qquad (1)$$

where a, b, and τ are positive real numbers; s is any real number; and Φ_1 is the degenerate hypergeometric function of two variables (Gradshteyn and Ryzhik, 1965, 9.261).

This g is a ratio of power series, and can be computed quite rapidly for a given tuple (a, b, τ, s) and a given Z. It leads to a large class of admissible estimators with a wide range of possible behavior. In particular, it includes many estimators that exhibit robustness to large values of Z; many estimators that offer significant risk reduction near $Z = 0$; and many that do both. This class generalizes the form noted by Maruyama (1999), which contains the positive-part James–Stein estimator as a limiting (improper) case.

2.2 Hypergeometric inverted-beta priors

The connection with multiple testing is as follows. Recall that under the alternative model, β_i is conditionally normal with variance λ_i^2. Our approach is to work with the transformed variable $\kappa_i = 1/(1 + \lambda_i^2)$, and to define the following prior for κ_i. Suppressing subscripts for the moment,

$$p(\kappa) = C^{-1}\kappa^{a-1}\ (1 - \kappa)^{b-1}\ \left\{\frac{1}{\tau^2} + \left(1 - \frac{1}{\tau^2}\right)\kappa\right\}^{-1} \exp(-s\kappa)\,, \qquad (2)$$

where $a, b, \tau > 0$ and $s \in \mathbb{R}$, and where C_1 is a constant of proportionality. We denote the hypergeometric-beta prior on the κ scale by $\kappa \sim \mathrm{HB}(a, b, \tau, s)$.

The normalizing constant,

$$C = \int_0^1 \kappa^{a-1}\ (1 - \kappa)^{b-1}\ \left\{\frac{1}{\tau^2} + \left(1 - \frac{1}{\tau^2}\right)\kappa\right\}^{-1} \exp(-s\kappa)\ \mathrm{d}\kappa\,, \qquad (3)$$

can be computed using hypergeometric series. In Appendix B we give details of this computation, which yields

$$C = e^{-s}\ \mathrm{Be}(a,b)\ \Phi_1(b,1,a+b,s,1-1/\tau^2)\,, \tag{4}$$

where Φ_1 is the degenerate hypergeometric function of two variables (Gradshteyn and Ryzhik, 1965, 9.261). This function can be calculated accurately and rapidly by transforming it into a convergent series of ${}_2F_1$ functions (§9.2 of Gradshteyn and Ryzhik, 1965; Gordy, 1998), making evaluation of (4) quite fast for most allowable choices of the parameters.

The implied density for λ_i^2 takes the form

$$p(\lambda^2) = C^{-1}(\lambda^2)^{b-1}\ (\lambda^2+1)^{-(a+b)}\ \exp\left\{-\frac{s}{1+\lambda^2}\right\}\ \left\{\tau^2+\frac{1-\tau^2}{1+\lambda^2}\right\}^{-1}\,. \tag{5}$$

This is a generalization of the inverted-beta distribution, also known as Pearson's Type VI distribution. Indeed, it reduces to an inverted beta in the special case where $s=0, \tau=1$, in which case $a\lambda^2/b$ will follow an $F(2b,2a)$ density.

The hypergeometric inverted-beta family contains many well-known sub-families of priors for κ. These include the beta distribution, the generalized beta distribution (McDonald and Xu, 1995), and the Gauss hypergeometric distribution (Armero and Bayarri, 1994). The family is itself contained in the class of compound confluent hypergeometric distributions (Gordy, 1998), which has two extra parameters that are not relevant in this context. These various related families are why we call (5) the hypergeometric inverted-beta prior. The transformed density on the κ scale resembles a beta distribution, and we call this family the hypergeometric-beta (HB) prior.

The family in (2) has one major advantage over other similar priors: there exist easily computable expressions for the posterior mean $\mathrm{E}(\beta_i \mid y_i)$ and the marginal density $m_1(y_i) = \int \mathrm{N}(y_i \mid \beta_i, \sigma^2)\ p(\beta_i)\ \mathrm{d}\beta_i$ under the hypothesis that $\beta_i \neq 0$. We derive these expressions in the Appendix.

2.3 Shrinkage profiles

We now turn to the specification of the four hyperparameters, and to the different "local shrinkage profiles" that are accessible through different choices of these parameters.

All normal scale-mixtures have an implied shrinkage profile $p(\kappa_i)$, which describes the amount of shrinkage toward the origin that is expected *a priori*. The prior's behavior near $\kappa_i = 0$ controls the tail weight of the marginal prior for β_i, while the behavior near $\kappa_i = 1$ controls the strength of shrinkage near zero.

Table 1 lists four common priors, while Figure 2 plots the implied shrinkage profiles for two of these: the double-exponential and Cauchy priors. Contrast these shrinkage profiles with the wide range of shapes that accessible through the hypergeometric inverted-beta density, some of which are shown in Figure 3.

One important special case of the hypergeometric inverted-beta family is the Strawderman prior (Strawderman, 1971), which corresponds to $a = 1/2$, $b = 1$, $s = 0$, and

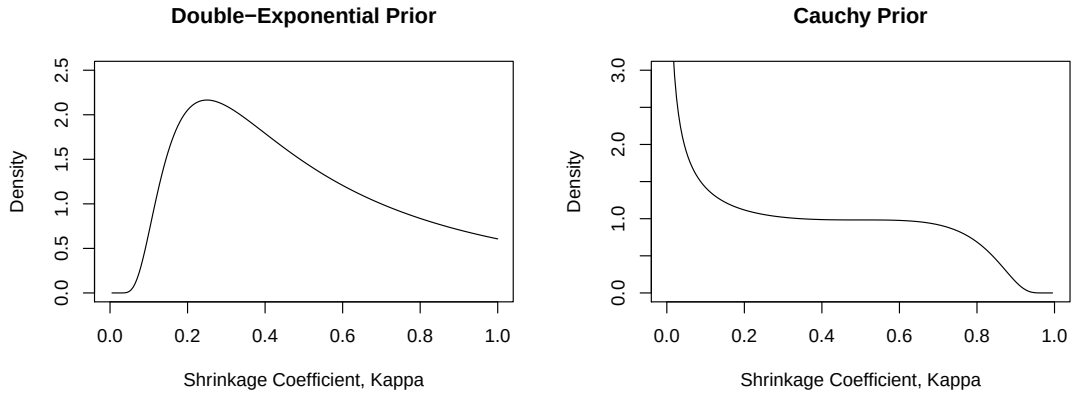

Figure 2: Implied shrinkage profiles for double-exponential and Cauchy priors.

Table 1: Priors for λ_i and κ_i associated with some common local shrinkage rules. Densities are given up to constant terms.

Prior for β_i	Prior for λ_i	Prior for κ_i
Double-exponential	$\lambda_i \ \exp\{\lambda_i^2/2\}$	$\kappa_i^{-2} e^{-\frac{1}{2\kappa_i}}$
Cauchy	$\lambda_i^{-2} \ \exp(-1/2\lambda_i^2)$	$\kappa_i^{-\frac{1}{2}}(1-\kappa_i)^{-\frac{3}{2}} e^{-\frac{\kappa_i}{2(1-\kappa_i)}}$
Strawderman–Berger	$\lambda_i \ (1+\lambda_i^2)^{-3/2}$	$\kappa_i^{-\frac{1}{2}}$
Horseshoe	$(1+\lambda_i^2)^{-1}$	$\kappa_i^{-1/2}(1-\kappa_i)^{-1/2}$

$\tau = 1$. Another special case is the half-Cauchy prior on the scale factor λ, studied by Gelman (2006) and Carvalho et al. (2010). This corresponds to $a = b = 1/2$, $s = 0$, and $\tau = 1$. Yet a third special case is the uniform-shrinkage prior, where $a = b = 1$, $s = 0$, and $\tau = 1$. All of these can be seen in the upper-left pane of Figure 3.

Clearly (2) can lead to many standard-looking shapes that are similar to other normal scale mixtures. Yet it can also produce a wide variety of other densities that are inaccessible through other standard families. We now describe the role of each hyperparameter, recalling that more probability near $\kappa = 1$ means more aggressive shrinkage.

First, τ is a global scaling factor, with larger values leading to larger marginal variance in β. To see this, suppose that all components of $\boldsymbol{\beta}$ have a common variance component in addition to their idiosyncratic ones: $(y_i \mid \beta_i) \sim \mathrm{N}(\beta_i, \sigma^2)$ and $\beta_i \sim \mathrm{N}(0, \sigma^2\tau^2\lambda_i^2)$. The form involving τ in (2) arises from the special case of assuming a half-Cauchy prior for each λ_i, as in the horseshoe prior of Carvalho et al. (2010). The generalization of the scaled half-Cauchy prior to arbitrary a, b, and s then arises quite naturally on the κ

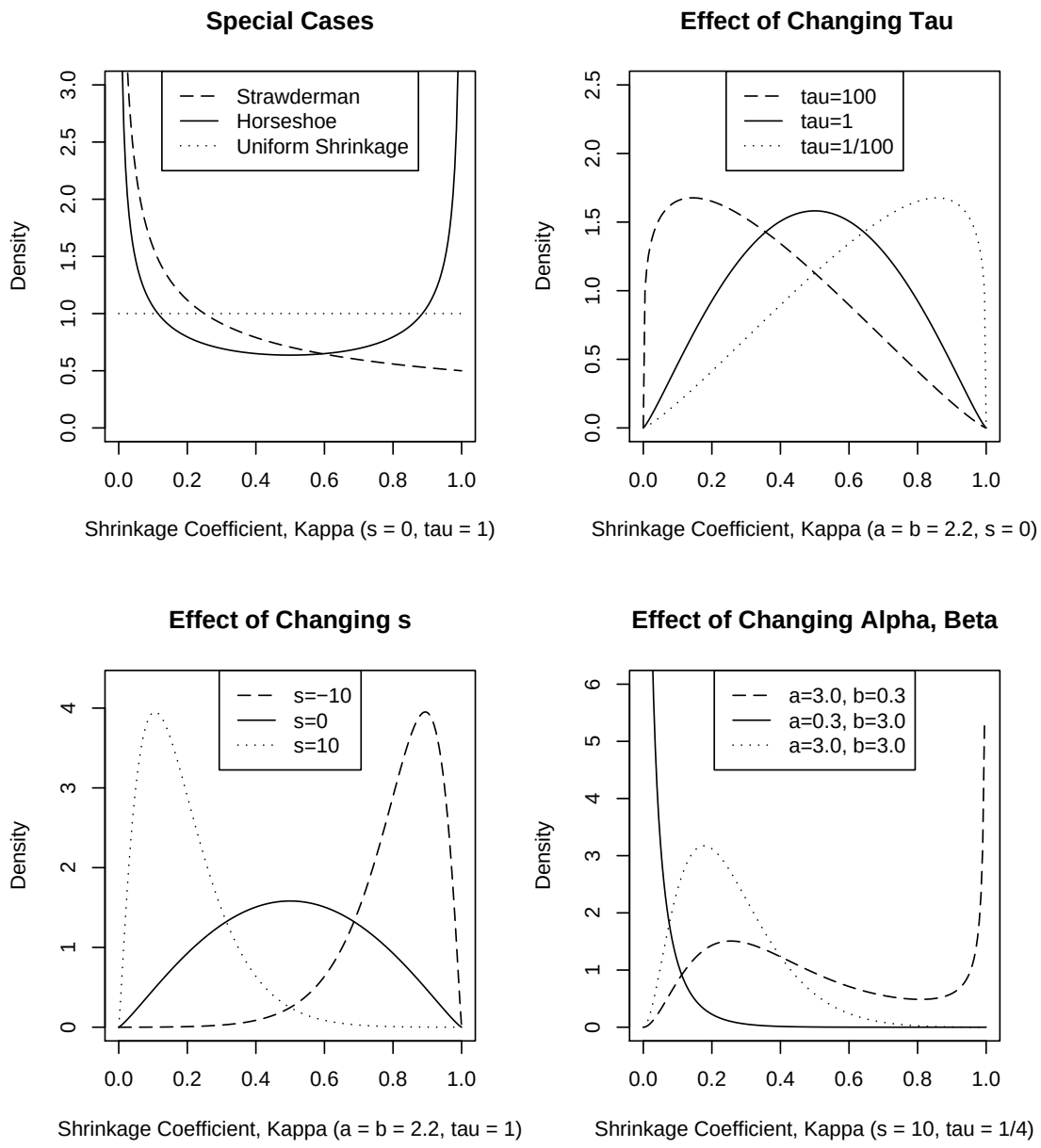

Figure 3: Effect of changing the four parameters (a, b, s, τ) on the density for the shrinkage coefficient κ.

scale. Shifting τ up and down causes the shrinkage profile to be shifted left and right, respectively, controlling the overall aggressiveness of shrinkage.

The parameters a and b are analogous to those of beta distribution, to which (2) reduces when $\tau = 1$ and $s = 0$. Smaller values of a encourage heavier tails in $\pi(\beta)$, with $a = 1/2$, for example, yielding Cauchy-like tails. Smaller values of b encourage $p(\beta)$ to have more mass near the origin, and eventually to become unbounded; $b = 1/2$ yields, for example, $p(\beta) \approx \log(1 + 1/\beta^2)$ near 0.

Finally, s is a second global scaling factor, though with a different effect than τ on the shape of the density. This parameter has an interpretation as a "prior sum of squares," with the caveat that it can also be negative.

The scale parameters τ and s do not control the behavior of $\pi(\lambda)$ at 0 and ∞. Specifically, $\pi(\lambda)$ behaves like λ_i^{2b-1} near the origin, and like $\lambda_i^{-(2a+1)}$ in the upper tail. Since $\pi(\beta)$ has the same polynomial rate of decay as $\pi(\lambda)$, a can be chosen to reflect the desired tail weight of $\pi(\beta)$.

2.4 The score function and overshrinkage of exceptional observations

We recall the following theorem from Carvalho et al. (2010).

Theorem 1. *Let $p(|y - \beta|)$ be the likelihood, and suppose that $p(\beta)$ is a mean-zero scale mixture of normals: $(\beta \mid \lambda) \sim N(0, \lambda^2)$, with λ having proper prior $p(\lambda)$. Assume further that the likelihood and $p(\beta)$ are such that the marginal density $m(y) < \infty$ for all y. Define the following three pseudo-densities, which may be improper:*

$$\begin{aligned} m^\star(y) &= \int_{\mathbb{R}} p(|y - \beta|)\ p^\star(\beta)\ d\beta\,, \\ p^\star(\beta) &= \int_{\mathbb{R}^+} p(\beta \mid \lambda)\ p^\star(\lambda)\ d\lambda\,, \\ p^\star(\lambda) &= \lambda^2 p(\lambda)\,. \end{aligned}$$

Then

$$\begin{aligned} E(\beta \mid y) &= \frac{m^\star(y)}{m(y)} \frac{d}{dy} \log m^\star(y) \\ &= \frac{1}{m(y)} \frac{d}{dy} m^\star(y)\,. \end{aligned} \tag{6}$$

Versions of this representation theorem appear in Masreliez (1975), Polson (1991), and Pericchi and Smith (1992). Theorem 1 relaxes a specific regularity condition having to do with the boundedness of $p(\beta)$, and extends the usual result to situations where $p(\beta)$ is a scale mixture of normals with proper mixing density and finite marginal $m(y)$.

The theorem characterizes the behavior of an estimator in the presence of large signals. Specifically, it says that we can achieve "inherent Bayesian robustness" by choosing a prior for β such that the derivative of the log predictive density is bounded as a function of y. Ideally, of course, this bound should converge to 0 for large $|y|$, will lead to $\mathrm{E}(\theta \mid y) \approx y$

for large $|y|$. This will avoid the overshrinkage of exceptional observations—clearly an important goal in large-scale simultaneous testing problems.

It is easy to verify, using the results of the previous subsection, that normal scale mixtures with hypergeometric inverted-beta mixing distributions satisfy the property of tail robustness. This helps to explain their good performance in high-dimensional settings.

2.5 The effect of shared shrinkage parameters

The hypergeometric inverted-beta prior allows a combination of global and local shrinkage that can be both flexible and robust. Figure 4 shows how a very small value of τ, encouraging strong global shrinkage, can be reinforced by a small observation ($y = 1.0$), and yet be almost completely overruled by a large observation ($y = 4.0$). Meanwhile, the marked bimodality for an intermediate observation such as $y = 2.5$ reflects uncertainty about whether such an observation corresponds to signal or noise, with the posterior mean for β averaging over both possibilities.

This example demonstrates that global shrinkage through τ can be very effective at squelching noise in high-dimensional problems. It is crucial, however, that τ be estimated from the data, and that the prior for κ_i grow sufficiently fast near 0 in order to allow κ_i to escape the strong "gravitational pull" of a small τ when y_i is large (as in this example when $y_i/\sigma = 4$). We recommend setting $a = 1/2$ in sparse problems involving a normal likelihood; see Carvalho et al. (2010) for further discussion. In situations with heavier-tailed sampling models, it may be appropriate to choose a smaller value of a.

When $1 - 1/\tau^2$ is very close to 1 (or when $1 - \tau^2$ is very close to 1 for $\tau < 1$), the Φ_1 functions may become slow to evaluate due to the slow convergence of the series representations given in the appendix. In our experience, the issue becomes practically significant in a serial computing environment only when τ^2 is larger than 1000 or smaller than 1/1000. Additionally, global shrinkage can take place through s rather than τ (with τ being set equal to 1). Then $\kappa_i \sim \mathrm{HB}(a, b, \tau = 1, s)$, and so

$$(\kappa_i \mid y_i) \sim \mathrm{HB}(a + 1/2, b, \tau = 1, s + y_i^2/2\sigma^2)\,.$$

Figure 5 shows that global shrinkage through s can produce results quite similar to global shrinkage through τ.

3 Multiple testing with heavy-tailed priors

Hypergeometric inverted-beta scale mixtures of normals are an especially useful class of priors for building discrete mixture models for β_i, due to the existence of simple expressions for moments and marginals under the hypothesis that β_i is nonzero:

$$\begin{aligned} (\beta_i \mid \kappa_i) &\sim w \cdot \mathrm{N}(0, \kappa^{-1} - 1) + (1 - w) \cdot \delta_0 \qquad (7) \\ \kappa_i &\sim \mathrm{HB}(a, b, \tau, s)\,, \qquad (8) \end{aligned}$$

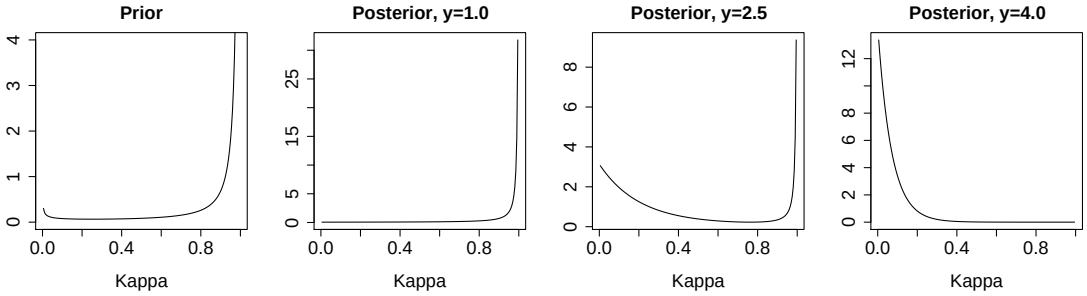

Figure 4: The left pane shows the prior for κ when $\tau = 1/15$, $s = 0$, and $a = b = 1/2$, reflecting a prior bias for strong shrinkage. The next three panes show the different posteriors for κ upon observing a single data point: $y = 1.0$, $y = 2.5$, or $y = 4.0$, respectively.

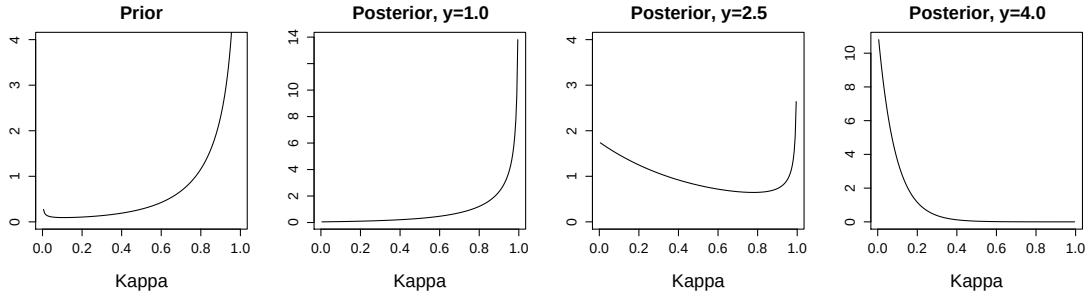

Figure 5: The left pane shows the prior for κ when $\tau = 1$, $a = b = 1/2$, and $s = -4$. The next three panes show the different posteriors for κ upon observing a single data point: $y = 1.0$, $y = 2.5$, or $y = 4.0$, respectively.

where δ_0 indicates a degenerate distribution at 0. The posterior mean under this model is a natural estimator for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$, since it averages over uncertainty about whether each component is zero or nonzero.

We conducted two simulation studies comparing the mean-squared error performance of our estimators with the procedure from Johnstone and Silverman (2004), where β_i is estimated by the posterior median under a mixture of a point mass zero and a double-exponential (Laplace) prior. We also keep track of the number of false positives generated by each procedure.

Each of the two studies involved estimating signals from a different signal class. In all cases the dimension of the location vector was $p = 1000$.

Experiment 1: Fixed coefficients of common size and varying sparsity levels. Table 2 summarizes an experiment involving 12 configurations of different sparsity patterns (5, 50, and 100 nonzero means) and different scales (all nonzero means equal to 3, 4, 5, or 7).

Experiment 2: Random t_3-distributed coefficients with varying sparsity levels. Table 3 summarizes an experiment in which the nonzero means were randomly drawn from a heavy-tailed t distribution with 5 degrees of freedom and scale parameter c. We investigated 12 configurations of different sparsity patterns (20, 50, 200, and 500 nonzero means) and different scales ($c = 0.5, 1, 2$).

Tables 2 and 3 show the average sum of squared errors in estimating $\boldsymbol{\beta}$ over 100 independent data sets. Also shown are the average number of false positives declared by the two procedures in each case, and the average false-discovery rate. For the Johnstone/Silverman procedure, a false positive occurs when the posterior median of β_i is nonzero, but the actual value is zero. For the Bayesian procedure using the hypergeometric–inverted-beta prior, a false positive occurs when the posterior inclusion probability for β_i is greater than 50% and β_i is actually zero. This threshold reflects a 0–1 loss function that penalizes false positives and false negatives equally, regardless of size. A full decision-theoretic analysis incorporating more realistic loss functions would yield a different, data-adaptive threshold, but would only complicate the analysis slightly.

For the hypergeometric inverted-beta prior, we set $s = 0$, while w and τ were estimated by importance sampling. For priors, we assumed that $\tau \sim \mathrm{C}^+(0, \sigma)$, and that $w \sim \mathrm{Unif}(0, 1)$.

In Experiment 1, we used a range of values for a and b. The best overall choice seemed to be $a = 1/2$, $b = 1/2$, and so we focused solely on this choice in Experiment 2. Indeed, although certain alternative choices produced improvements in specific situations, we found $a = b = 1/2$ to be a good all-purpose option because of its blend of good performance in estimation and testing.

Overall, when squared error in estimation is used to decide between procedures, our preferred Bayes procedure with $a = b = 1/2$ wins slightly on Experiment 2, while the empirical-Bayes thresholding procedure wins slightly on Experiment 1. We attribute these differences to the relative tail weight of the two priors. The double-exponential prior has tails that are heavier than the Gaussian likelihood, but not as heavy as those of the

Table 2: Experiment 1, fixed coefficients. SSE: sum of squared errors in the estimate of the $\boldsymbol{\beta}$ sequence. FP: false positive declarations in the estimate of $\boldsymbol{\beta}$ sequence. FDR: realized false-discovery rate. Laplace: posterior median estimator from the empirical Bayes procedure of Johnstone and Silverman (2004). The numbers in parentheses indicate, in order, the choices of a and b the HIB model.

		Number nonzero out of 1000 means											
		5				50				100			
	Value	3	4	5	7	3	4	5	7	3	4	5	7
SSE	Laplace	35.1	32.8	17.9	8.5	210.5	150.8	99.7	71.9	331.1	248.3	177.5	142.9
	$(1, 2)$	35.4	31.9	17.9	10.3	205.4	157.7	116.7	90.6	334.6	268.2	213.2	180.4
	$(1, 1)$	35.0	31.3	18.5	11.1	200.5	161.9	124.7	95.3	329.1	280.8	229.3	188.5
	$(1, 0.5)$	34.7	31.0	19.6	12.2	199.6	170.7	135.3	100.6	335.2	302.2	248.1	196.3
	$(0.5, 2)$	37.9	36.8	18.3	7.3	242.6	167.3	104.0	70.8	395.3	272.8	182.8	145.7
	$(0.5, 1)$	37.6	36.3	18.1	7.6	234.9	164.1	105.0	72.6	379.5	268.8	186.4	148.9
	$(0.5, 0.5)$	37.4	35.7	17.9	7.9	227.5	161.1	106.2	74.2	363.6	266.2	190.9	151.9
FP	Laplace	0.8	1.0	0.8	0.4	16.1	11.3	7.6	4.2	53.3	28.7	17	8.9
	$(1, 2)$	0.2	0.3	0.6	0.5	4.0	6.9	6.6	5.5	12.2	18.2	17.2	13.2
	$(1, 1)$	0.2	0.4	0.7	0.5	6.4	10.2	9.4	6.9	23.7	34.0	29.2	18.7
	$(1, 0.5)$	0.3	0.6	0.8	0.7	13.5	21.1	16.9	9.8	153.5	199.8	90.0	33.4
	$(0.5, 2)$	0.1	0.1	0.1	0.2	1.1	2.5	2.2	2.2	2.9	5.5	5.4	5.1
	$(0.5, 1)$	0.1	0.1	0.2	0.2	1.4	3.0	2.7	2.5	3.7	7.1	6.7	5.9
	$(0.5, 0.5)$	0.1	0.1	0.2	0.2	1.7	3.7	3.1	2.8	5.5	9.5	8.6	6.8
FDR	Laplace	0.2	0.2	0.1	0.1	0.3	0.2	0.1	0.1	0.4	0.2	0.1	0.1
	$(1, 2)$	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.1
	$(1, 1)$	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.1	0.2	0.3	0.2	0.2
	$(1, .5)$	0.2	0.1	0.1	0.1	0.3	0.3	0.2	0.2	0.6	0.6	0.5	0.2
	$(0.5, 2)$	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.1	0.1	0.0
	$(0.5, 1)$	0.1	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.1	0.1	0.1	0.1
	$(0.5, 0.5)$	0.1	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

Table 3: Experiment 2, random coefficients. The HIB prior set $a = 1/2$, $b = 1/2$.

		Number nonzero											
		50			100			200			500		
	Scale c	0.5	1	2	0.5	1	2	0.5	1	2	0.5	1	2
SSE	HIB	8.3	16.0	55.4	28.8	53.2	125	90.2	235	336	181	391	604
	Laplace	8.6	16.1	60.4	29.5	57.3	136	93.1	250	370	180	394	646
FP	HIB	0.0	0.0	0.2	0.0	0.2	0.5	0.1	0.8	3.3	0.1	0.9	10.8
	Laplace	0.4	3.7	1.1	2.3	1.6	3.2	23.5	34	19.1	138	134	71.5

hypergeometric inverted-beta priors we studied. This difference in tail weight becomes much more significant in the experiment with random coefficients, since draws from a t_3 density produce some very large signals—much larger than signals of size 7 in the "fixed coefficients" study. In Experiment 2, however, the heavier-tailed priors are wasting some of their mass in areas of the parameter space far from the origin. Since these areas are predestined to be unimportant by the particular choices of fixed signals, it is no surprise that a lighter-tailed prior such as the double-exponential will yield superior results. Similarly, when the coefficients are slightly larger, as in the t_3 signals from Experiment 2, the heavier-tailed prior will outperform.

But when the measuring stick is the false-positive rate, the fully Bayes procedure with smaller values of a and b wins. It produces far fewer false positives across the board, along with lower false-discovery rates (suggesting that it is not merely more conservative across the board in declaring an observation to be a signal). It therefore seems like the more robust choice. For situations when estimation is the goal, its performance is roughly comparable to the existing Johnstone/Silverman procedure. Yet for situations when testing is the goal, the Bayes procedure appears more trustworthy.

4 Testing for superior historical performance

4.1 Summary of preliminary results

We ran the proposed multiple-testing method on the cohort of firms for which at least 10 years of past data were available. This initial sieve left 11,298 firms, each with somewhere between 10 and 43 annual observations. Based on the simulation results above, we are reporting results for $a = b = 1/2$.

Of these, 424 firms had posterior inclusion probabilities larger than 90%, indicating moderate to high confidence that they have systematically outperformed their peer groups. The top 10 firms, all with posterior inclusion probabilities higher than 98%, are list described in Table 4, along with the reason for dropping out of the data base (if applicable). Of these 10 firms, 8 seemed to outperform their peer group, while 2 seemed to underperform. The first non-American firm on the list is British–American Tobacco, incorporated in (of all places) Malaysia, which ranks 11th by estimated posterior inclusion probability.

The historical trajectories for these 10 firms can be seen in Figure 6. Two are large drug companies; the rest come from a variety of different industries. All but four—Wyeth, Merck, Tambrands, and WD-40—are likely unknown to the average consumer.

4.2 Comparison with the popular literature on corporate success

It is interesting to compare these results to the conclusions of certain well-known books that purport to explain corporate success. We took a small, nonscientific sample of these books, in an attempt to gauge whether the results from the multiple-testing model correspond to widely held notions about successful firms. Table 5 briefly describes these books, and indicates whether the basis for selecting the study cohort was qualitative or quantitative in nature. The books were chosen in conjuction with a group of senior management

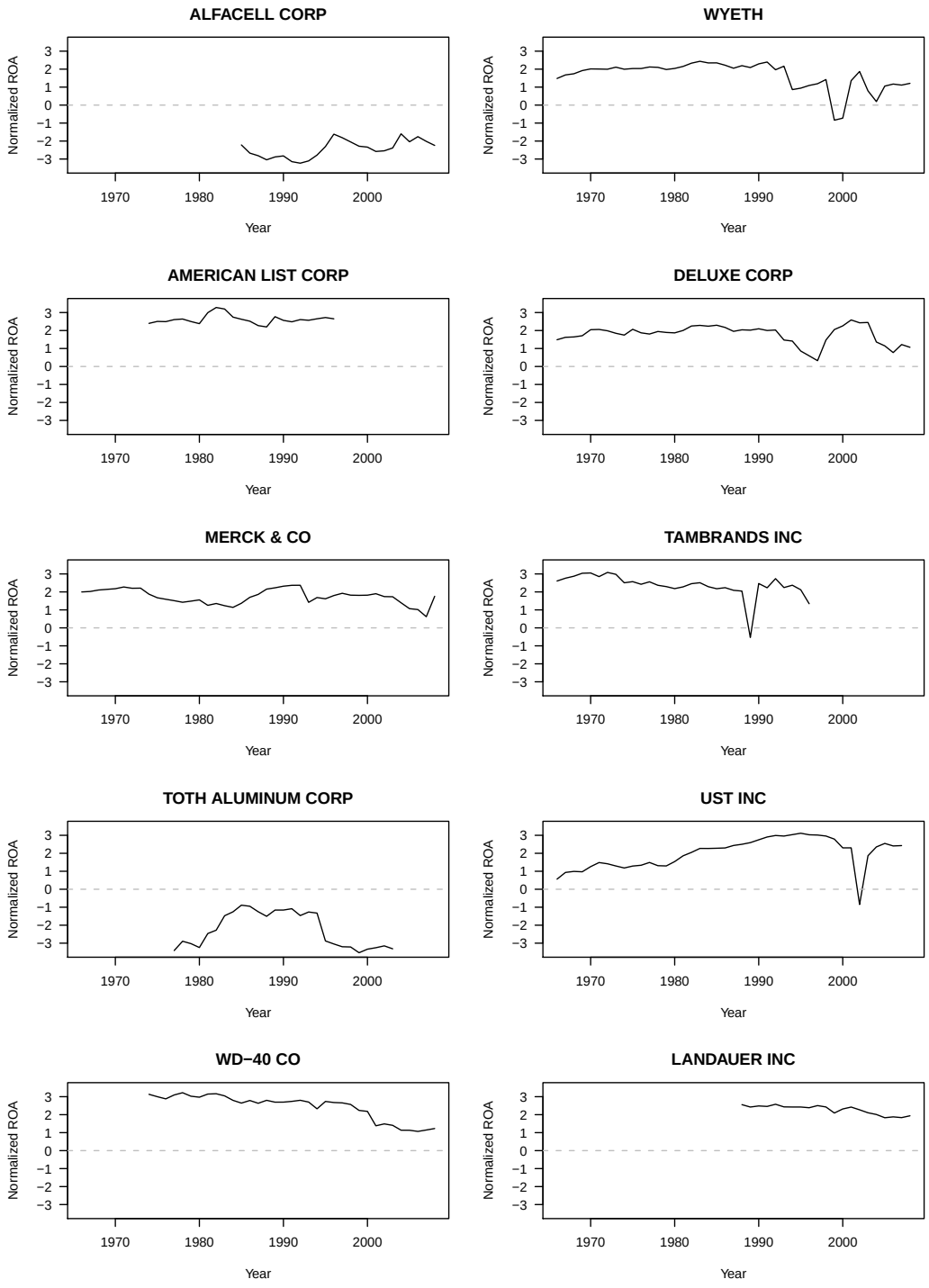

Figure 6: The performance trajectories for the ten firms with the highest posterior probabilities of having a nonzero mean.

Table 4: Ten firms with the highest posterior probabilities of having a nonzero mean.

Company	Description	Books
Alfacell Corporation	A biotechnology firm specializing in RNA-based technologies.	—
Wyeth	Large drug company; recently bought out by Pfizer.	—
American List Corp	Maintains lists of addresses for bulk mailing. Bought out in 1997.	—
Deluxe Corp	Specializes in financial and logistical services for small businesses.	—
Tambrands	Manufactures personal hygiene products. Bought out in 1997.	—
Toth Aluminum	Developed technology for producing aluminum. Defunct.	—
UST	A tobacco holding company. Bought out in 2009.	—
WD-40	Manufactures the eponymous anticorrosive and lubricating agent.	—
Landauer	Specializes in services relating to radiation safety.	—
Merck	Another large drug company.	BTL, ISE

consultants at Deloitte Consulting, who judged the list to be fairly representative of the popular literature.

These books follow a common recipe: start with a group of companies; identify the "successful" ones; look for patterns in their behavior; and abstract those behaviors into a small set of principles that can tell others how to run their businesses better.

None of these studies, however, make a serious attempt to verify statistically that the selected companies have done anything special when compared with a suitable reference population. This opens up the possibility that they have been studying companies that were lucky, rather than great—the precise null hypothesis considered in this paper.

Indeed, serious discrepencies emerged between the popular literature and the conclusions of the multiple-testing procedure considered here. Across the nine books considered, there were 209 distinct firms that were used as case studies—some positive, some negative—and that also appeared in our cohort of firms with 10 or more years of data. Of the top ten firms flagged in the previous section, only one was mentioned in any of the 9 books: Merck, a case study in Built to Last (BTL) and In Search of Excellence (ISE).

Of the 209 firms mentioned as case studies, 27 of them (13%) were flagged by the multiple-testing model as having greater than 95% probability of having a nonzero mean. Of the firms not mentioned in any book, 397 (or 3.5%) of them were thus flagged. This is a noticeable difference, to be sure, and one borne out by a simple chi-squared test: the hypothesis of independence between the two states of "flagged by a book" and "flagged by the testing procedure" is summarily rejected ($p < 0.001$). But the degree of overlap is perhaps not as much as might be expected. The sheer number of firms that were held up in these nine books as examples worthy of emulation—and yet failed to be flagged by the multiple-testing procedure, even under the generous assumption of year-on-year independence—is worrisome.

Table 5: The popular books selected for comparison.

Title	Published	Selection method	Basis
Good to Great	2001	Companies from 1965–1981 selected on the basis of shareholder return	Quantitative
Built to Last	1994	Companies founded before 1950 that met certain success criteria	Qualitative
In Search of Excellence	1982	Based off surveys of executives at author-selected firms	Qualitative
Competitive Strategy	1980	Author selected examples to support theory; method unclear	Qualitative
Hidden Values	2000	Author selected examples to support theory; method unclear	Qualitative
Blueprint to a Billion	2006	Based off time to achieve $1 billion in revenue after initial public offering	Quantitative
What Really Works	2003	Based on correspondence with consultant-identified "top management practices"	Qualitative
Stall Points	2008	Based off revenue-growth stalls and/if revenue growth recovery	Quantitative
Blue Ocean Strategy	2005	Author selected examples to support theory; method unclear	Qualitative

4.3 Robustness to assumption of independence

As a test statistic, we have used the composite z-score $z_i = \bar{z}_i\sqrt{n_i}$ for each firm. This implicitly assumes that a company's ROA result in year t is independent of results from previous years.

To relax this assumption, we apply the same analysis to a new quantity $z_i(\alpha)$, a rescaled version of the z-score that is expressed as a function of $\alpha \in [0, 1]$:

$$z_i(\alpha) = \bar{z}_i n_i^{(1-\alpha)/2}.$$

When $\alpha = 0$, these are identical to the original z-scores, corresponding to n_i independent data points. Increasing α smoothly from 0 to 1 implies an increasing degree of dependence between successive observations, and a corresponding deflation in the perceived impressiveness of a firm's performance. In the extreme case where $\alpha = 1$, we are treating the composite z-score as though it has arisen from only a single data point, representing a case of extreme residual autocorrelation. By varying α and recomputing results, we can determine the assumed "effective sample size" where significant results evaporate. Further analyses of the individual time series can help quantify what their individual effective sample sizes seem to be.

For $\alpha = 0.5$, for example, only 8 firms have posterior inclusion probabilities greater than 95%, and only 1226 have inclusions probabilities greater than 50%. This reduces the effective sample size to 3.2 at the low end (10 years of data), and 6.5 at the high end (43 years of data). By the time $\alpha = 0.75$—a choice that cuts 43 years of data down to an effective sample size of only 2.6 years—only 3 firms cross the 50% threshold: American

List Corp, Tambrands, and WD-40.

5 Final Remarks

We have developed a Bayesian multiple-testing procedure based upon a heavy-tailed prior for the nonzero means. These priors form an interesting, novel class of normal variance mixtures, the hypergeometric–inverted-beta class. Overall, the procedure has the nice theoretical property of a redescending score function under the alternative model, and seems to perform as well as, or better than, existing gold-standard methods. Moreover, it allows relevant Bayesian and frequentist summaries to be computed with minimal computational fuss. This property arises from the simple, known form of the marginal distribution $m(y)$.

We have applied the method to a large data set on historical corporate performance, and compared the results of our analysis to some popular books that deal with the same subject. Another recent study along these lines was done by Henderson et al. (2009), who reach similar conclusions. By our measure, these books appear to be studying a sample wherein the majority of firms have performance profiles that are statistically indistinguishable from luck. Meanwhile, there are a few hundred firms (out of a group of over 10,000) whose performance is at least suggestive of a sustained advantage, and yet were not considered in these high-profile case studies.

In other words, these books seem to have failed to study the companies most likely to be better than their peers, and may very well be studying pure noise. This has obvious negative consequences for the confidence one can place in these books, whose advice appears uncomfortably haphazard in light of the evidence presented here.

A Expressions for moments and marginals

Throughout this section, we suppress conditioning on β_i's nonzero status. Under our hypergeometric inverted-beta model, the joint distribution for y_i and κ_i takes the form

$$p(y_i, \kappa_i) \propto \kappa_i^{a'-1} \, (1-\kappa_i)^{b-1} \, \left\{ \frac{1}{\tau^2} + \left(1 - \frac{1}{\tau^2}\right) \kappa_i \right\}^{-1} e^{-\kappa_i s'} \,,$$

where now $s' = s + y_i^2/(2\sigma^2)$ and $a' = a + 1/2$.

The moment-generating function of (2) is easily shown to be

$$M(t) = e^t \, \frac{\Phi_1(b, 1, a+b, s-t, 1-1/\tau^2)}{\Phi_1(b, 1, a+b, s, 1-1/\tau^2)} \,.$$

See, for example, Gordy (1998). Expanding Φ_1 as a sum of ${}_1F_1$ functions and using the differentiation rules given in Chapter 15 of Abramowitz and Stegun (1964) yields

$$\mathrm{E}(\kappa^n \mid \mathbf{y}, \sigma^2) = \frac{(a')_n}{(a'+b)_n} \frac{\Phi_1(b, 1, a'+b+n, s', 1-1/\tau^2)}{\Phi_1(b, 1, a'+b, s', 1-1/\tau^2)} \,. \tag{9}$$

Using (9), we get

$$\mathrm{E}(\beta_i \mid y_i) = \left\{1 - \frac{a'}{a'+b} \frac{\Phi_1(b, 1, a'+b+1, s', 1-1/\tau^2)}{\Phi_1(b, 1, a'+b, s', 1-1/\tau^2)}\right\} y\,. \tag{10}$$

And by the law of total variance,

$$\begin{aligned} \mathrm{Var}(\beta_i \mid y_i) &= \mathrm{E}\{\mathrm{Var}(\beta_i \mid y_i, \kappa_i)\} + \mathrm{Var}\{\mathrm{E}(\beta_i \mid y_i, \kappa_i)\} \\ &= \sigma^2\{1 - \mathrm{E}(\kappa_i \mid y_i)\} + y^2\ \mathrm{Var}(\kappa_i \mid y_i)\,, \end{aligned} \tag{11}$$

will all other posterior moments for β_i following in turn.

There is also a tractable expression for the marginal likelihood of the data:

$$m(y_i) = C_1^{-1} \int_0^1 \kappa_i^{a'-1}\ (1-\kappa_i)^{b-1}\ \left\{\frac{1}{\tau^2} + \left(1 - \frac{1}{\tau^2}\right)\kappa_i\right\}^{-1} e^{-\kappa_i s'}\ \mathrm{d}\kappa_i\,, \tag{12}$$

where again $s' = s + y_i^2/(2\sigma^2)$ and $a' = a + 1/2$. This integral is in the same family as (3), and so by the same series of arguments we obtain

$$m(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}}\ \exp\left(-\frac{y_i^2}{2\sigma^2}\right)\ \frac{\mathrm{Be}(a', b)}{\mathrm{Be}(a, b)}\ \frac{\Phi_1(b, 1, a'+b, s', 1-1/\tau^2)}{\Phi_1(b, 1, a+b, s, 1-1/\tau^2)}\,. \tag{13}$$

B Details of hypergeometric inverted-beta integrals

Theorem 2. *The hypergeometric inverted-beta density is proper for all $a, b, \tau > 0$ and $s \in \mathbb{R}$.*

Proof. The normalizing constant in (2) is

$$C = \int_0^1 \kappa^{\alpha-1}\ (1-\kappa)^{\beta-1}\ \left\{\frac{1}{\tau^2} + \left(1 - \frac{1}{\tau^2}\right)\kappa\right\}^{-1} \exp(-s\kappa)\ \mathrm{d}\kappa\,. \tag{14}$$

Let $\eta = 1 - \kappa$. Using the identity that $e^x = \sum_{m=0}^{\infty} x^m/m!$, we obtain

$$C = e^{-s} \sum_{m=0}^{\infty} \left[\frac{s^m}{m!} \int_0^1 \eta^{\beta+m-1}(1-\eta)^{\alpha-1}\{1 - (1-1/\tau^2)\eta\}^{-1}\ \mathrm{d}\eta\right]\,.$$

Using properties of the hypergeometric function ${}_2F_1$ (Abramowitz and Stegun, 1964, §15.1.1 and §15.3.1), this becomes, after some straightforward algebra,

$$C = e^{-s}\ \mathrm{Be}(\alpha, \beta)\ \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{(\beta)_{m+n}}{(\alpha+\beta)_{m+n}\ m!\ n!}\ s^m\ (1 - 1/\tau^2)^m\,, \tag{15}$$

where $\mathrm{Be}(\cdot, \cdot)$ is the beta function and $(a)_n$ is the rising factorial. Appendix C of Gordy (1998) proves that, for all $\alpha > 0$, $\beta > 0$, and $1/\tau^2 > 0$, the nested series in (15) converges

to a positive real number, yielding

$$C = e^{-s}\ \mathrm{Be}(\alpha, \beta)\ \Phi_1(\beta, 1, \alpha + \beta, s, 1 - 1/\tau^2)\,, \tag{16}$$

where Φ_1 is the degenerate hypergeometric function of two variables (Gradshteyn and Ryzhik, 1965, 9.261). □

The Φ_1 function can be written as a double hypergeometric series,

$$\Phi_1(\alpha, \beta; \gamma; x, y) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{(\alpha)_{m+n} (\beta)_n}{(\gamma)_{m+n} m! n!}\ y^n\ x^m\,, \tag{17}$$

where $(c)_n$ is the rising factorial. We use three different representations of $\Phi_1(\alpha, \beta, \gamma, x, y)$ for handling different combinations of arguments, all from Gordy (1998). When $0 \leq y < 1$ and $x \geq 0$,

$$\Phi_1(\alpha, \beta, \gamma, x, y) = \sum_{n=0}^{\infty} \frac{(\alpha)_n}{(\gamma)_n} \frac{x^n}{n!}\ {}_2F_1(\beta, \alpha + n; \gamma + n; y)\,. \tag{18}$$

When $0 \leq y < 1$ and $x < 0$,

$$\Phi_1(\alpha, \beta, \gamma, x, y) = e^x\ \sum_{n=0}^{\infty} \frac{(\gamma - \alpha)_n}{(\gamma)_n} \frac{(-x)^n}{n!}\ {}_2F_1(\beta, \alpha; \gamma + n; y)\,. \tag{19}$$

Finally, when $y < 0$,

$$\Phi_1(\alpha, \beta, \gamma, x, y) = e^x\ (1 - y)^{-\beta}\ \Phi_1(\tilde{\alpha}, \beta, \gamma, -x, \tilde{y})\,, \tag{20}$$

where $\tilde{\alpha} = \gamma - \alpha$ and $\tilde{y} = y/(y - 1)$. Then either (18) or (19) may be used to evaluate the righthand side of (20), depending on the sign of x.

Alternative representations for Φ_1 involving ${}_1F_1$ functions are also available. In our experience, however, these take longer to converge than those given above.

References

F. Abramovich, Y. Benjamini, D. Donoho, and I. Johnstone. Adapting to unknown sparsity by controlling the false-discovery rate. *The Annals of Statistics*, 34(2):584–653, 2006.

M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, volume 55 of *Applied Mathematics Series*. National Bureau of Standards, Washington, DC, 1964. Reprinted in paperback by Dover (1974); on-line at `http://www.math.sfu.ca/$\sim$cbm/aands/`.

C. Armero and M. Bayarri. Prior assessments for predictions in queues. *The Statistician*, 43:139–53, 1994.

Y. Benjamini and Y. Hochberg. Controlling the false-discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300, 1995.

J. O. Berger. A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *The Annals of Statistics*, 8(4):716–761, 1980.

M. Bogdan, A. Chakrabarti, and J. K. Ghosh. Optimal rules for multiple testing and sparse multiple regression. Technical Report I-18/08/P-003, Wroclaw University of Technology, 2008a.

M. Bogdan, J. K. Ghosh, and S. T. Tokdar. A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*, volume 1, pages 211–30. Institute of Mathematical Statistics, 2008b.

C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–80, 2010.

D. B. Dahl and M. A. Newton. Multiple hypothesis testing by clustering treatment effects. *Journal of the American Statistical Association*, 102(478):517–26, 2007.

K.-A. Do, P. Muller, and F. Tang. A Bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society, Series C*, 54(3):627–44, 2005.

B. Efron. Microarrays, empirical Bayes and the two-groups model (with discussion). *Statistical Science*, 1(23):1–22, 2008.

D. Fourdrinier, W. Strawderman, and M. T. Wells. On the construction of Bayes minimax estimators. *The Annals of Statistics*, 26(2):660–71, 1998.

A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Anal.*, 1(3):515–33, 2006.

M. B. Gordy. A generalization of generalized beta distributions. In *Finance and Economics Discussion Series*. Board of Governors of the Federal Reserve System, 1998.

I. Gradshteyn and I. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, 1965.

R. Gramacy and H. K. Lee. Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103(483):1119–30, 2008.

J. Griffin and P. Brown. Alternative prior distributions for variable selection with very many more variables than observations. Technical report, University of Warwick, 2005.

A. D. Henderson, M. E. Raynor, and M. Ahmed. How long must a firm be great to rule out luck? benchmarking sustained superior performance without being fooled by randomness. In *The Academy of Management Proceedings*, 2009.

W. James and C. Stein. Estimation with quadratic loss. In *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, volume 1, pages 361–79, 1961.

H. Jeffreys. *Theory of Probability*. Oxford University Press, 3rd edition, 1961.

I. Johnstone and B. W. Silverman. Needles and straw in haystacks: Empirical-Bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594–1649, 2004.

F. Liang, R. Paulo, G. Molina, M. Clyde, and J. Berger. Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103:410–23, 2008.

Y. Maruyama. Improving on the James–Stein estimator. *Statistics and Decisions*, 14: 137–40, 1999.

C. Masreliez. Approximate non-Gaussian filtering with linear state and observation relations. *IEEE. Trans. Autom. Control*, 1975.

J. B. McDonald and Y. J. Xu. A generalization of the beta distribution with applications. *Journal of Econometrics*, 66:133–52, 1995.

P. Muller, G. Parmigiani, and K. Rice. FDR and Bayesian multiple comparisons rules. In *Proceedings of the 8th Valencia World Meeting on Bayesian Statistics*. Oxford University Press, 2006.

J. Park and J. Ghosh. A guided random walk through some high dimensional problems. *Sankhya, Series A*, 72(1):81–100, 2010.

L. R. Pericchi and A. Smith. Exact and approximate posterior moments for a normal location parameter. *Journal of the Royal Statistical Society (Series B)*, 54(3):793–804, 1992.

N. G. Polson. A representation of the posterior mean for a location model. *Biometrika*, 78:426–30, 1991.

J. G. Scott and J. O. Berger. An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136(7):2144–2162, 2006.

C. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9:1135–51, 1981.

W. Strawderman. Proper Bayes minimax estimators of the multivariate normal mean. *The Annals of Statistics*, 42:385–8, 1971.

A. Zellner and A. Siow. Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia*, pages 585–603, 1980.